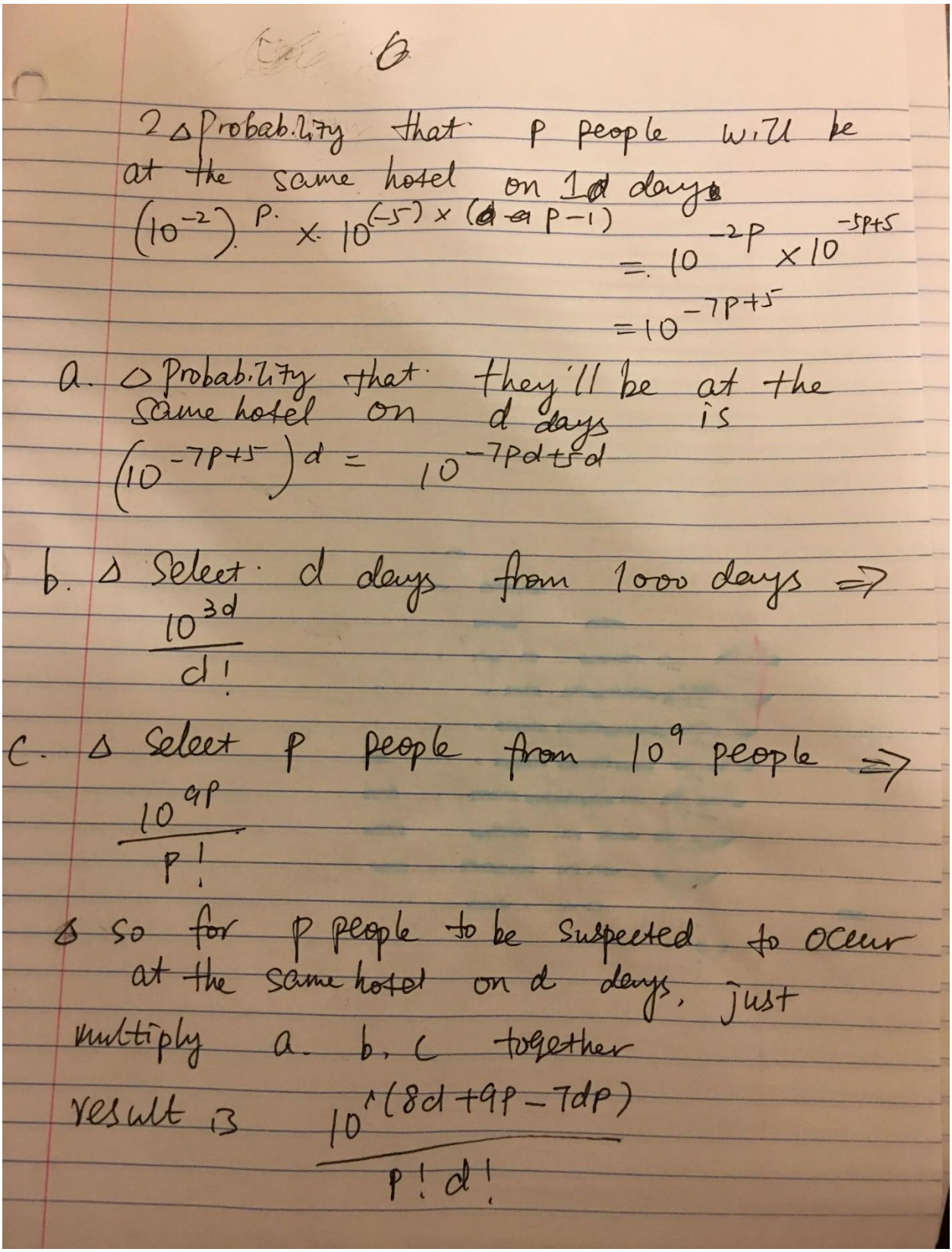
Michael Wang – wustl key: michaelwang

Terry Lyu ----- wustl key: terrylu

HW1

1. Big Data Properties
    a. Log Data:
        i. Amount of records: There is an entry in a log file every second if not more, so the amount of data points must be huge.
        ii. Infinity: Likely the log files are being generated faster than we can process them and are useful in that we can filter them to look at what happened in the past, but most likely can't process in real time.
        iii. Structure: The log data seems to be fairly structured with the date, time, ip address, location, another measurement of time, and some 4 digit integer afterwards. There's definitely a pattern among the entries
    b. Wikipedia articles:
        i. Amount of records: Wikipedia has a vast amount of entries.
        ii. Complexities: Often Wikipedia articles reference other articles so there is a very complex relationship between all the different records.
    c. Database of chemical compounds:
        i. Amount of records: I believe we talked in class how there are thousands of chemical compounds added every day to the ZINC database so the amount of records is likely very large.
        ii. Structure: the data seems to be formatted with headers such as "bonds" and "bond IDs".
        iii. Labels: The "costly to observe property" seems like it could be considered a label since it is difficult to obtain information.
    d. Differences in datasets?
        i. The log data likely has far more records since it is a data dump of what is occurring on many different computers/servers. It also has a much faster stream that the Wikipedia articles. Additionally, it seems very well structured, whereas the Wikipedia articles are just a collection of sentences.



2.
3. Unreasonable Effectiveness of Data:

a. Labeled/annotated data has been marked up and given attributes by humans. Unlabeled data on the other hand, has not been described and had notes written about it, though there is far more of it. An example of each could be a corpus vs a random html web page.
b. The data-based approach is to create an algorithm and model that can interpret unlabeled data because there is such a vast amount of unlabeled data, and you should follow the data, not create a model for data that doesn't exist.
c. Unlabeled data is difficult to work with because you need to interpret it. You aren't given any sort of starting point and it is generally unstructured.

4. Bonus Problem.
    a. I thought the reading was a bit over my head so that was frustrating. The first question was definitely reasonable, and the second question has me stuck. Overall I think this was a moderate homework, probably a bit difficult for the first one. The second question definitely caught me off guard.